III. Technological Challenges in the Genome Project

Improving Sequencing and Finding Genes

Bob Moyzis: Let's turn to the problem of improving sequencing technology. In order to carry out the vision Maynard gave—sequencing not only the human but also the mouse genome and all the human cDNAs and so forth—we need at least a hundredfold improvement in sequencing efficiency.

David Cox: There's another way to think about this problem. Suppose we focus not on sequencing the whole human genome, but on a more manageable goal—namely, finding out how to determine the sequence of 2 million base pairs of DNA accurately and rapidly. That achievement would have an absolutely revolutionary impact on human biology because it would provide an ideal tool for finding disease genes.

The search for the single dominant gene that causes Huntington's Disease illustrates my point. In 1983, a DNA marker that was very tightly linked to Huntington's Disease was identified by Jim Gusella and his colleagues. It's now 1992 and the Huntington's Disease gene has yet to be identified. The research has been narrowed down to a region of DNA 2.5 million base pairs long. Yeast artificial chromosomes and other mapping techniques have allowed most of that region to be cloned, so we actually have the DNA in hand. A number of groups across the world are dedicated to the search, but we still don't have the Huntington's Disease gene. Why not? Well, in that 2.5 million base pairs of DNA, there are probably fifty different genes. And how do we find out which one is the Huntington's gene? There's just no easy solution.

The problem of finding a single base change in 2 million base pairs of DNA is going to be the standard problem in finding disease genes.

Right now the approach is first to identify all of the genes in that region, say by hybridization to cDNAs, and then look for abnormalities in those genes. If the disease gene contains a DNA rearrangement, it's easy to identify. Or perhaps the messenger RNAs from the disease gene are different in size or amount from those of the normal gene. If we compare the messenger RNA of each of the fifty genes from a Huntington's patient and from a normal individual, we might be able to identify the disease gene.

But chances are that the Huntington's gene won't contain a DNA rearrangement and won't change the size or the amount of the messenger RNA. So even if all fifty genes are identified, we will probably have to sequence all fifty genes from a Huntington's patient first, and then from unaffected individuals to identify changes present only in Huntington's patients but never in normal individuals. That will be the proof that you have found the Huntington's mutation. In fact, that exact strategy was used to prove that the cystic-fibrosis gene was indeed the disease-causing gene.

Suppose instead that you could sequence the whole region known to contain the Huntington's gene and find out what base changes are present only in Huntington's patients and never in normal patients. Then you could identify the disease gene immediately, and you wouldn't have to mess around finding all the genes in the region.

The problem of finding a single base change in 2 million base pairs of DNA is going to be the standard problem in finding disease genes. So if we had a way of sequencing 2 million base pairs accurately and rapidly, it would completely revolutionize how we went about finding human disease genes, and it would cut down the amount of work by at least a factor of ten. After sequencing the region, we could use PCR-based assays to examine very quickly the DNA from 100 normal individuals and thereby distinguish harmless polymorphisms from the disease-causing mutation. But we can't carry out this approach because present sequencing technology is simply too remedial.

Bob Moyzis: Whether we're searching for disease genes or wanting to sequence the whole genome, sequencing technology is currently not up to the job. However, incremental changes in current technology during the next few years are likely to increase the rate of sequencing to between a hundred thousand and a million nucleotides per day. Thin-gel technology, pioneered by Lloyd Smith and others, has been demonstrated to yield a tenfold improvement in throughput simply by increasing the voltage used to separate the DNA molecules. Further, parallel processing of samples using robotics or other more exotic techniques, such as flow cytometry, is being pursued. Advances in primer walking, such as those being developed at Brookhaven National Laboratory by Bill Studier, also look promising for the near term. We would need a major breakthrough to process a billion base pairs per day, but a million base pairs per day will be within reach at many laboratories in the next few years. As David Cox has said, at that rate most of the interesting goals of the Human Genome Project can be achieved.

David Galas: I agree that refinements in current technology will yield the tenfold to one hundredfold improvement that Bob is talking about. At that rate the bottleneck will not be sequencing but rather front-end preparation and back-end analysis. The back end, which includes entering short stretches of sequence, 300 to 800 bases long, that come off the sequencing machine into the database, assembling those sequences into long, contiguous sequences, checking for errors, and so on, needs great improvement [see "DNA Sequencing"].

It's time for the DOE to do production-line or large-scale sequencing so we can find the hang-ups in those areas and address them. Sequencing technology itself should be seen as one module among many in this process, a module that can be changed as better technology comes along.

Lee Hood: I'm glad to hear you say that because a major output of the Genome Project is going to be DNA sequence data. Until now the DOE has done a super job of supporting the development of radically new sequencing technologies, which may—or may not—lead to a hundredfold or a thousandfold increases in output, but we also need to do the systems integration required for large-scale sequencing projects with present technologies. That's the only way to learn the requirements for setting up production-line, large-scale, fully-automated technologies of the kind that will be needed to sequence the entire human genome.

Over the next ten years, we're hoping to get at least a hundredfold increase in sequencing throughput because that's what it will take to carry out the genome initiative. If we succeed, then I don't think academics will do the sequencing; it will be industry.

David Galas: The DOE is sponsoring some sequencing of model organisms now, and we're thinking seriously about setting up pilot sequencing projects, the principle goal of which would be to understand the bottlenecks in production-line sequencing and to identify the places where new technology would really help.

So far we have only begun to scratch the surface of problems associated with sequence assembly and error checking. We haven't had enough data to work on. Later, when sequencing costs and efficiencies, including front and back ends, improve by a factor of at least ten, it probably would be appropriate to start sequencing large, selected regions of the human genome.

Lee Hood: We should also encourage industry to get involved in such projects. Over the next ten years, we're hoping to get at least a hundredfold increase in sequencing throughput because that's what it will take to carry out the genome initiative. If we succeed, then I don't think academics will do the sequencing; it will be industry. Sequencing companies will get subcontracts from the government for large-scale sequencing. Industry needs to get involved now, so that when the technology is ready for high-throughput sequencing, they'll have skilled people to carry it out.

If we set up this large-scale sequencing effort now, I think we could produce a million base pairs of accurate, or finished, sequence per person, per year. We're still learning how to do this and various problems slow us down. The production of the DNA fragments for sequencing is not trivial. Each fragment must be sequenced five or six times to reduce sequencing errors. The assembly of long sequences from overlapping, short sequences is not fully automated, and the clones are not always faithful copies of the genome. To do large-scale sequencing we have to figure out how to make all these steps move faster in a reliable and integrated system.

Bob Moyzis: Determining the correct sequence would seem to be very important, but we know that a single sequencing run can produce an error rate as high as 1 in 100. That means that the disease-gene hunts described by David Cox would be very inefficient. The sequencing of a 2-million-base region would produce 200,000 errors. That's an awful lot of data to check. Lee, how do you currently deal with errors?

Lee Hood: We deal with the errors in two ways. First we're doing the shotgun sequencing method, that is, we're picking many clones at random and sequencing them. Those clones overlap each other, so on average, we're sequencing each stretch of DNA between six and seven times. That gives us an error rate of perhaps 1 in 5000. Second, for each cloned fragment, we sequence about 15 percent of the cloning vector. Since the vector sequence is known, we determine the error rate for each run through the machine. Some runs have more errors because the chemical reactions used to prepare the DNA for the machines may have worked poorly and so forth. To my mind, the error rate in sequencing is not an insurmountable difficulty. It's true that an error rate of 1 in 100,000 is going to cost a lot of money, but if we can live with an error rate of 1 in 1000 or 1 in 5000, we'll be in good shape.

Many of the errors in sequencing are due to problems at the front end of the process. Cloning artifacts, such as deletions, for example, are not uncommon. Those artifacts are likely to be much more frequent in human DNA and mouse DNA because those genomes contain an abundance of repetitive sequences. Such sequences are probably a substrate for nonhomologous recombination, which, if it occurs during the cloning process, can create new sequences not present in the genomic DNA.

So any DNA that has lots of repetitive sequences is intrinsically less stable than DNA lacking repetitive sequences. We could use better cloning systems for minimizing those artifacts, but short of that, we'll probably develop much better ways of checking clones to make sure they match their germ-line counterparts before we start analyzing them. Perhaps the hybridization-based technologies will be important both in mapping clones and in checking sequenced DNA for errors.

> *Ten years ago, if a good graduate student produced 12,000 base pairs of finished sequence in a year, that was considered very good. Today a machine can do 12,000 base pairs of rough sequence each day.*

Bob Moyzis: Lee, what are you doing on the front and back ends of sequencing?

Lee Hood: At Caltech we haven't done much with the front-end problems because Applied Biosystems is developing a robot for doing the PCR and the standard sequencing reactions in a format that's consistent with placing the reaction products directly into a fluorescence sequencing machine. On the back end, we're working together with LOBE on two major projects. First, we're developing a laboratory management system to keep track of all the details that are a part of sequencing—where the fragments came from, how they've been treated, what time they were run on the machines, and so forth. Second, we're working on computer programs for assembling a long sequence from randomly generated short sequences. They still need a lot of work. With the fluorescence sequencing machine, a computer program reads the order of the nucleotide bases directly from the sequencing gel and puts question marks in positions of ambiguity. Someone must look at the data and make decisions regarding those question marks. In the future, we should have better programs for *calling* the sequences. To do large-scale sequencing, we will have to automate this whole process in a way that requires a minimum of manual intervention.

Bob Moyzis: Earlier I voiced my optimism that these problems will be solved. I know you share that optimism.

Lee Hood: We need to remind ourselves of the progress we've made over the last ten years. Ten years ago, if a good graduate student produced 12,000 base pairs of finished sequence in a year, that was considered very good. Today, a machine can do 12,000 base pairs of rough sequence each day. Thus we've had an increase of several orders of magnitude in throughput. I think the front- and back-end problems are more straightforward and are going to be solved. The problem of getting good robots to prepare the reaction mixtures is technically less demanding than figuring out how to improve DNA sequencing throughput by two orders of magnitude.

David Galas: Given the uncertainty in meeting those demands, we have to plan on some large-scale sequencing using present-day, conventional technologies. But the new technologies are coming along, and there are two kinds. Those that push the present methods include multiplex sequencing, automated multiplex sequencing, capillary-gel electrophoresis, and automatic detection systems. And we can expect those developments to yield a tenfold improvement—maybe even more.

Then, there are three or four radically new technologies that could change things dramatically. One is the Los Alamos single-molecule-detection method [see "Rapid DNA Sequencing Based on Single-Molecule Detection"]. That's gotten to the point where they can actually detect single molecules.

A lot of progress is also being made on hybridization sequencing. Even if it doesn't work for precise sequencing, it'll work for gathering partial sequences of a lot of DNA extremely rapidly. The idea is to place huge numbers of short sequences, eight to ten bases long, on a little chip and determine which of those hybridize to the long fragment being sequenced. In its ultimate form, these hybridizations yield the full sequence, but even partial sequence information will be helpful for mapping and for finding homologous regions. Right now there's too much noise in the system, so the hybridization signals aren't clean. But those problems are being worked on, and I would say that the hybridization method is neck-and-neck with the Los Alamos single-molecule-detection scheme.

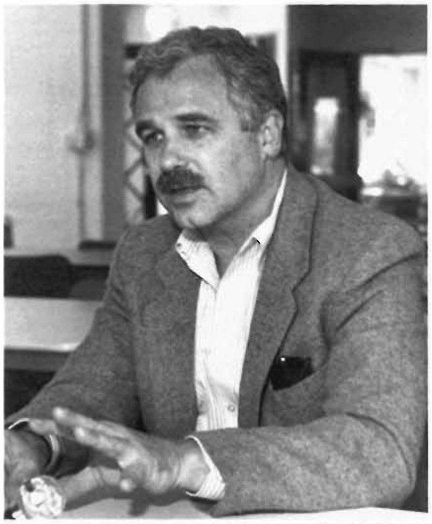
The other new sequencing method uses mass spectrometry. You start with the set of fragments produced by normal sequencing reactions. Remember, those are a set of nested fragments that increase in length stepwise, that is, one base at a time. You arrange to place a single charge on each of these, and then you use a laser to blast the stuff off a little plate into a vacuum. Because all the fragments are charged equally, you can use a device to separate them by mass and get the whole ladder of fragments laid out in a single measurement. It takes only a few milliseconds. If that method works with the required accuracy, you can read the sequence instantly. It requires measuring the mass of these fragments to one part in a few thousand so that you can determine which base is at each place in the sequence. That's a radically new idea.

Lee Hood: The center at Caltech is focused on improving sequencing technologies, and there, we're taking two approaches. One is to implement a better design of the contemporary automated fluorescence sequencing machine by using better lasers, thin gels, pulse-field gel electrophoresis, and the like.

The second approach is to explore whether mass spectrometry can really be used for sequencing. As David Galas explained, the idea is to measure the mass of each of the fragments generated from the standard sequencing reactions. You can either measure the masses of the fragments from the four different dideoxy reaction mixtures, or if the resolution is higher, you can measure the masses of all the fragments from a combined mixture. For the latter, you have to have a resolution that can distinguish single-nucleotide additions.

David Galas: With three or four of these completely new ideas under development, my guess is that sooner or later one of them is going to work well enough for practical application and will revolutionize sequencing. My bet is that we're going to have some of these working within five to ten years, which is about when we were hoping to start doing some serious large-scale sequencing.

If one of these methods works, we'll be able to do what David Cox was talking about. We could sequence the chromosomes of an affected individual as well as the chromosomes of unaffected individuals, and we would be able to identify immediately what mutations were responsible for a given condition.



David Galas

With three or four of these completely new ideas under development, my guess is that sooner or later one of them is going to work well enough for practical application and will revolutionize sequencing. My bet is that we are going to have some of these working within five to ten years, which is about when we were hoping to start doing some serious large-scale sequencing.

Technology Development—an interdisciplinary challenge

Bob Moyzis: It's clear from the problems we're facing in mapping and sequencing that this project requires technological development in every area.

David Botstein: We're weak enough in technology that we really ought to invite people from other disciplines: chemistry, physics, robotics, and the like, to think about it. We hope that this issue of *Los Alamos Science* will reach people who can come out of the woodwork to help us. And I think it's really important to distinguish between what really helps and what doesn't help. We don't need a lot of physicists to turn themselves into biologists. But we do need physicists who have enough interest in the biology and enough patience to understand what the technical problems are.

I'll give you two examples from my own life. Around 1975 when I was at MIT, we were taking electron-microscope pictures of DNA. DNA looks like little worms with kinks in them. There's a lot of information in those little worms and we were using a map measurer to figure out how long the contour lengths were. We went to the computer group, which had a PDP-9, and we said, "Can you do this for us automatically?" And they said, "Get lost, kid, it's trivial." So finally, I got a Master's student and bribed him to look at this problem. He took it to his boss and they came back a month later and said, "Not only is it not trivial, but it's impossible. We can't do it." Of course what he really meant was that he didn't think he was going to get anything out of solving the problem—it wouldn't get him tenure.

Today there is still no automatic equipment to make that measurement. It still can't be done. But we have to find a way to collaborate because I think that both sciences would benefit greatly.

Bob Moyzis: The cultural problem is very real. On the one hand biologists think of biological solutions to the problems. And one of the beauties of biology is that you can manipulate a bug to do your work for you, so there's a resistance to tapping into the physical-science community. It's only recently that low-key robotics has even entered biology. Maybe that's because molecular biologists think it's good for the soul to do these repetitive tasks.

Biologists think of biological solutions to the problems. And one of the beauties of biology is that you can manipulate a bug to do your work for you, so there's a resistance to tapping into the physical-science community.

On the other hand, if the physical scientists think they're being used to solve a trivial problem, they are never going to get interested. They have to feel that their contributions to the goals of this project are exciting and worth doing.

David Botstein: Steve Chu is a laser physicist who has been working with DNA at Stanford. He has invented a contraption that can stretch out an individual piece of DNA and measure its length by how far it stretches before it breaks. That's the kind of thing that would be fun to do. But Steve is unusual in that his brother is the biologist who invented the CHEF gel. So it's a special case because they talk to each other.

Every manipulation that we do in the Genome Project is suboptimal. For example, when people take pictures of in-situ hybridizations, they use cooled ccd-array cameras that are probably three generations old. Physicists wouldn't dream of using one of those. They're probably piling up as junk in the basement of the CERN accelerator.

Bob Moyzis: Certainly the general problem of image analysis or pattern recognition needs better solutions. We're using very antiquated technology, for example, in analyzing our gels. In many areas of biology, we're swamped and would love to find a way to automatically extract data, enhance images, and look for patterns, be they linear or three-dimensional.

David Botstein: Part of our five-year plan is technology development, but right now we don't know who are the right people to talk to. We think that we have employment for at least the next ten or fifteen years for these interdisciplinary guys. But they don't exist. They literally don't exist.

Nancy Wexler: We are trying to create a new kind of interdisciplinary science with a leg in not just physics and biology, but in other disciplines as well. We need to appeal to young people who are just beginning their training and who are willing to be a little experimental. We need meetings to define the issues and the problems and to bring people from different disciplines together. Then we need a specialized training program.

Maynard Olson: The Genome Project clearly needs a strong engineering component, and maybe that's another reason Wally Gilbert says the Project isn't science. Basic scientists often look down on engineering, but most don't know much about it. Some of the most creative things done in the 20th century have been engineering advances. When the dust settles on this century, we'll look back on two great technological revolutions: one in computers, the other in DNA technology.

Computers are largely an engineering advance. Early on new theoretical ideas about managing digital information and advances in solid-state physics were critical, but the real surge in computing power came when creative engineers took over and built better and better computers. We're not talking about building a slightly better mousetrap; we're talking about creating compositions of matter whose behavior differs qualitatively from anything people a few years before would have thought possible. Computers are an open-ended technology where a factor of ten improvement in memory or processing speed sets the stage for another factor of ten. At any given stage in the technology, it's always the imagination of the users that is limiting, but they catch up remarkably quickly.

There is a real analogy here between computers and DNA. I suspect that creative engineering on this basically monotonous chemical will open up applications in biology as important as those opened up by modern computers. The underlying idea behind computers was that if one got extremely good at processing digital information, one could do an immense variety of things with it. Similarly, if we could analyze DNA—whether that means mapping, sequencing, or whatever—ten times better than we do now, it would yield tremendous opportunities for biological research and biomedical applications. When that happens, people won't be able to imagine working in the previous environment. What's more, the next factor of ten will have a similar impact.

Right now we're not working from this generic approach to DNA experimentation, but it will happen. I have my own ideas about how we might proceed, and I'm sure other ideas are out there. Such activities will not be a trivial mechanization of the present manual processes. It will mean taking a zero-based view of what we're trying to accomplish with DNA—and of the various physical tools that could be brought to bear on accomplishing those goals. That's the attitude we'll gradually grow into in DNA research. And I believe creative engineers will play a big role.

Lee Hood: We knew from the beginning that this project is about technology development, and to do that you need scientists who have interdisciplinary skills, who can talk to people, encourage new insights, and set up collaborations across different disciplines. These scientists are not easy to find. For the future, we need to establish training programs that cut across the different disciplines.

As far as getting things done now, we have to identify scientists who want to make a major commitment to the goals of the Project, either to produce highly informative genetic maps, or to make a physical map of a particular chromosome, or to do large-scale sequencing. Few scientists have

David Botstein

Part of our five-year plan is technology development, but right now we don't know who are the right people to talk to. We think that we have employment for at least the next ten or fifteen years for these interdisciplinary guys. But they don't exist. They literally don't exist.

Lee Hood

The national labs are set up to do interdisciplinary projects, but until recently, they haven't been that strong in biology, and this project must be directed by scientists who really understand biology.

made this commitment. But if people who are now making the appropriate commitments were funded in an appropriate fashion, more people would be encouraged to take on these larger tasks.

At Caltech, we have a very strong interdisciplinary program by virtue of our NSF-funded Science and Technology Center. We have groups working on nucleic acid chemistry, computational problems, genetic mapping and DNA diagnostics, and large-scale sequencing. They are all housed together and are an incredibly interactive group. And it's the close interaction that really makes things happen.

Bob Moyzis: Lee, you were the prime mover behind development of automated sequencing machines. Tell us a bit about that development.

Lee Hood: I've been involved in technology development throughout my career. I got my Ph.D. training in protein chemistry and then switched over into molecular biology. Soon after Gilbert and Sanger came out with their groundbreaking sequencing techniques, we started trying to develop an automated sequencing machine.

For about three years, we went about it in the wrong way. We essentially tried to develop a very clever way of reading the standard four-lane radioactive gels. But each lane of a gel has its own artifacts, which may put the bands in one lane ahead of the bands in another, or create zig-zags in the mass scale from one lane to the next. Those artifacts are due to temperature anisotropies, and so forth.

My view now is that four-lane sequence analysis has intrinsic difficulties in accuracy, whereas putting all four reaction mixtures in one lane allows the fragments from each mixture to be used as an internal standard against one another, so you get much more accurate sequence readings.

That is the approach we took in developing the automated fluorescent sequencing machine. Tim Hunkapillar first suggested the use of fluorescent tags on the DNA fragments produced by the enzymatic sequencing reactions. The tagged fragments are run down a single-lane gel past a laser, the laser causes the tags to fluoresce, and the color of the signal tells you which base was on the end of that fragment.

Lloyd Smith, a very good chemist from Stanford who joined our group in the early 1980s, developed those technologies. He'd had experience with lasers and was the right person at the right time. We also had a good organic chemist, Rob Kaiser, who could synthesize four different fluorescent compounds. So, putting together an interdisciplinary team of physical chemists, organic chemists, biologists, and then engineers, who could actually build the machine, was the key to making it work.

A lot of good universities are ideal places for interdisciplinary work because they have good departments in physics, computer science, engineering, and chemistry. Caltech is unusual because it is quite small, so it's easy for us to get to know people in different disciplines. That's much harder to do at the bigger universities.

On the other hand, the national labs are set up to do interdisciplinary projects, but until recently, they haven't been that strong in biology, and this project must be directed by scientists who really understand biology. For example, Bob Moyzis has contributed enormously to the genome center at Los Alamos.

Bob Moyzis: Thanks for the compliment. I have been somewhat frustrated by people from the physical sciences who seem interested in the mapping problem, in the physical reality of DNA, but who don't understand what the technical problems really are. A few years ago, mathematical types had a real obsession with modeling the best way to map the genome, and yet little of that theoretical work has had an impact on the experimental work.

Maynard Olson: That's because the modeling phase didn't pay adequate attention to experimental practicality. The mapping problem is dominated by the fact that the data aren't perfect, and a pristine model that assumes perfect data yields essentially no insight into the path that should be followed. So a purely theoretical approach to mapping problems won't help.

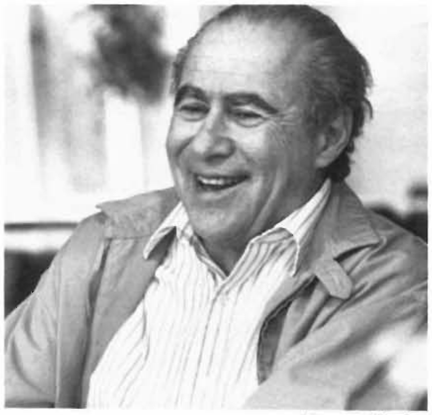
People in experimental physical chemistry, for example, have a better feel for the interplay between experiment and structure. The people who did the original molecular-beam experiments were very suspicious of pure theoreticians who wanted to take everything back to the wave equation. But they understood that there was quantization and that if they designed their detectors right, they could measure molecules in different quantum states, and they went on with the job. It also helped greatly that there were many investigators who were skilled in both theory and experiment. We do not have many people in biology with comparable breadth.

Norton Zinder: I've always had difficulty communicating with the theoretical physicists. Leo Szilard used to come to my lab suggesting experiments on DNA control and DNA synthesis that were meaningless because they were impossible to do. I spent four hours talking to the *great god of physics*, Neils Bohr, who supposedly had great ideas about biology, but I never understood what he was talking about. He could not relate to the experimental system I was trying desperately to describe to him. So the theoretical physicists probably won't be of much help. But people working on materials science do appreciate the complexities of biology and know how to think about experimental systems.

Maynard Olson: I am also concerned about the present generation of molecular-biology graduate students. Too many of them don't know much about either molecules or biology. What they know is how to manipulate DNA, to do Northern blots and Southern blots, and site-directed mutagenesis and so forth. However, this problem may be a transitory response to two decades during which these protocols largely defined molecular biology.

The brighter young molecular biologists are beginning to study developmental biology and pathology, for example, and to work with transgenic mice. They're looking at livers again. They're starting to learn some biology, and some are starting to learn a lot about molecules. Biophysics is enjoying a renaissance with nice work on protein folding and recognition of macromolecules by other macromolecules.

Another new front will be people working on genome mapping. Those mappers—or whatever they're to be called—are going to be people with different backgrounds, and they'll be more specialized. Molecular biology has just been through a gold-rush phase, a phase when the techniques were crude and the participants were jacks-of-all-trades. They did the genetics, they did the sequencing, they did protein chemistry, and they made a start at getting out the



Norton Zinder

I've always had difficulty communicating with the theoretical physicists. Leo Szilard used to come to my lab suggesting experiments on DNA control and DNA synthesis that were meaningless because they were impossible to do.

information in DNA. But, just as serious mining operations require assayists, surveyors, lawyers, mining engineers, and the like, if we are going to get out all the information contained in the genome, we need specialists in all the techniques related to DNA analysis.

David Cox: This field is in its formative stages, and it's the obligation of the scientific community to identify areas where technology development can really help. Then it's up to the Genome Project to put money into those areas. The scientists who sit back and criticize the Project but don't know what they want or don't come forth with suggestions are missing a great opportunity.

We need requests that are posed carefully. If you want more rapid ways of sequencing the human genome, then the question remains: What's rapid enough? But if you say, "I want to sequence 2 million base pairs of DNA in the next eight months, can you do it or can't you?" then it's a concrete job and the question has a concrete answer.

The Genome Project is designed to solve concrete problems. We need new technology, but we also need to put it into action. This country is grappling on many fronts with the issue of getting technology out to the people who can use it. For example, the United States has invented a lot of the basic devices that are used in the electronics industry, but those devices are not being marketed or manufactured in the U.S. They're being manufactured in other countries. The goals of the Genome Project are not just to invent things but also to manufacture and come through with the goods. Inventing technology doesn't do the deed. It's delivering that technology that counts, and the Genome Project will be successful only if it does both.

David Galas: David Cox is right. We need to deliver good products to the biomedical community. But we should not forget that this is a fundamental science project as much as it is a medical one. At any time in this project, we're going to have some defined goals that we're working towards, but I don't think we should consider the present five-year goals as sacrosanct, or fixed. After all, they were made up by guys thinking about the way things were two or more years ago. We'll probably change the five-year goals, and those changes will depend on the changing technology.

> *The Genome Project is forcing a bunch of researchers to cooperate and exchange information . . . that new way of working is going to change the sociology of how we do science. Rather than . . . working quietly in isolation and then giving a talk at a meeting maybe once or twice a year, many of us are learning a different way of doing projects, and I think it's all very healthy.*

Bob Moyzis: Watch it! Many of us drafted those five-year goals.

David Galas: Okay, let me give a really radical scenario. Let's suppose it turns out to be very easy to do the genetic and physical mapping for the mouse genome and extremely difficult to do it for the human genome. Then we ought to map everything on the mouse first and go back to humans later. The strategy we adopt will depend on how the technology works out.

This is an interesting time for biology. I think that most people don't realize how much the Human Genome Project is going to change the way we do biology. We're learning to take on huge tasks, and quite frankly, most of them are still above us. We are taking on tremendously broad goals, and we are realizing just how information-intensive this field is. We need new developments in automation, and we also need to interface with computers to the same extent that people in physics and chemistry do.

Five or ten years from now, I expect that the standard molecular-biology laboratory will be completely different from what it is today. There won't be any glassware. People will just have machines and computers. We will have automated the manipulations of DNA and animal cells, and we'll be able to go after fundamental biological problems with enormously powerful tools.

The Genome Project is forcing a bunch of researchers to cooperate and exchange information over computer networks, and that new way of working is going to change the sociology of how we do science. Rather than everyone going back to his or her lab, working quietly in isolation, and then giving a talk at a meeting maybe once or twice a year, many of us are learning a different way of doing projects, and I think it's all very healthy.

DNA Sequencing

An understanding of the structure, function, and evolutionary history of the human genome will require knowing its primary structure—the linear order of the 3 billion nucleotide base pairs composing the DNA molecules of the genome. Determining that sequence of base pairs is the long-term goal of the 15-year Human Genome Project. Both the merits and the technical feasibility of sequencing the entire human genome are discussed in Parts I and III of "Mapping the Genome." The bottom line is that sequencing technology is not yet up to the job.

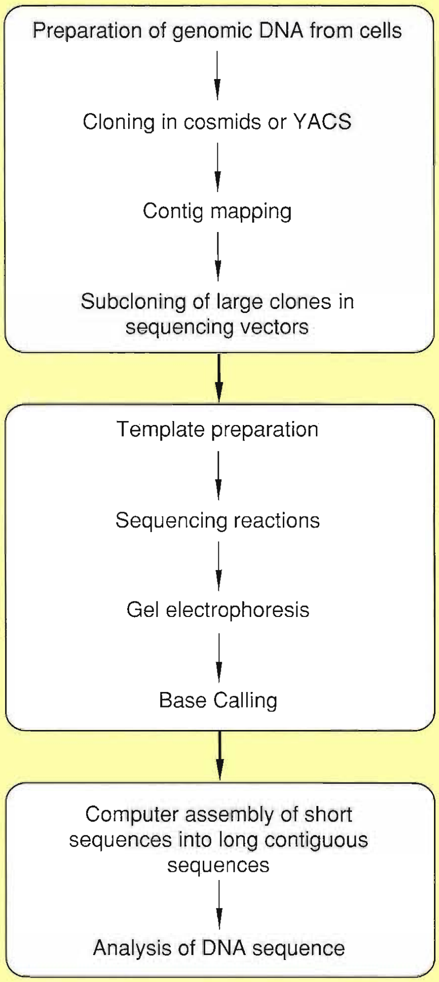
In 1990, when the plans for the Genome Project were being made, the estimated cost of sequencing was $2 to $5 per base. That is, a single person could produce between 20,000 and 50,000 bases of "finished" sequence per year. The term "finished" sequence implies the error rate is very low (the conservatives say an error rate of 1 base in 10^5 is acceptable, and the less conservative say 1 in 10^3 or 10^4). A low rate is achieved, in part, by sequencing a given region many times over. The planners agreed that the costs of sequencing must be substantially reduced and that the rate of producing finished sequence must increase by a factor of 100 to 1000 for sequencing the entire human genome to become an affordable and practical goal.

On the other hand, sequencing technology has been improving steadily for the past two decades. In the early 1970s one person would struggle to complete 100 bases of sequence in one year. Then two very similar techniques were developed—one by Allan Maxam and Walter Gilbert in the United States and the other by Fredrick Sanger and his coworkers in England—that made it possible for one person to sequence thousands of base pairs in a year. Those techniques, for which the inventors were jointly awarded the Nobel Prize, still form the basis of all current sequencing technologies. Both methods are described in greater detail below.

Between 1975 and the present, the number of base pairs of published sequence data grew from roughly 25,000 to almost 100 million. During that time longer and longer contiguous stretches of DNA have been sequenced. In 1991 the longest sequence to be completed was that of the cytomegalovirus genome, which is 229,354 base pairs. By 1992 a cooperative effort in Europe had sequenced an entire chromosome of yeast, chromosome III, which is 315,357 base pairs. And now efforts are underway to sequence million-base stretches of DNA. Accomplishing such large-scale sequencing projects is among the goals for the first five years of the Genome Project.

In order to achieve this goal, each step in the multi-stage DNA sequencing process must be streamlined and smoothly integrated. Figure 1 outlines all the steps involved in the sequencing of long, contiguous stretches of genomic DNA, DNA isolated from the genome. The initial steps include cloning large fragments of genomic DNA in YACs or cosmids and using those clones to construct a contig map for the regions to be sequenced. The contig map arranges the cloned fragments in the order and relative positions in which they appear along the genome. The cloning and mapping steps are described elsewhere in this issue (see "DNA Libraries" and "Physical Mapping").

Figure 1. Steps in Large-Scale Sequencing


Preparation of genomic DNA from cells
Cloning in cosmids or YACS
Contig mapping
Subcloning of large clones in sequencing vectors
Template preparation
Sequencing reactions
Gel electrophoresis
Base Calling
Computer assembly of short sequences into long contiguous sequences
Analysis of DNA sequence

To determine the DNA sequence of the mapped region, the large DNA insert in each of the large clones must be broken into smaller pieces of a size suitable for sequencing, and those small pieces must be cloned. This subcloning is often done in the cloning vector M13, a bacteriophage whose genome is a single-stranded DNA molecule. M13 accepts DNA inserts from 500 to 2000 base pairs in length, propagates in the host cell *E. coli*, and is particularly convenient for the Sanger method of sequencing. Each of the small clones is then sequenced.

As mentioned above, all sequencing technologies currently in use are based on the Sanger or the Maxam-Gilbert method, which were developed in 1977. Both methods determine the sequence of only one strand of a DNA molecule at a time, and both methods involve three basic steps. Below we mix and match certain technical details of each method to simplify the description of these three steps. The real methods are described in Figures 4 and 5.

Figure 2. Nested Set of Labeled Fragments for Simplified Example

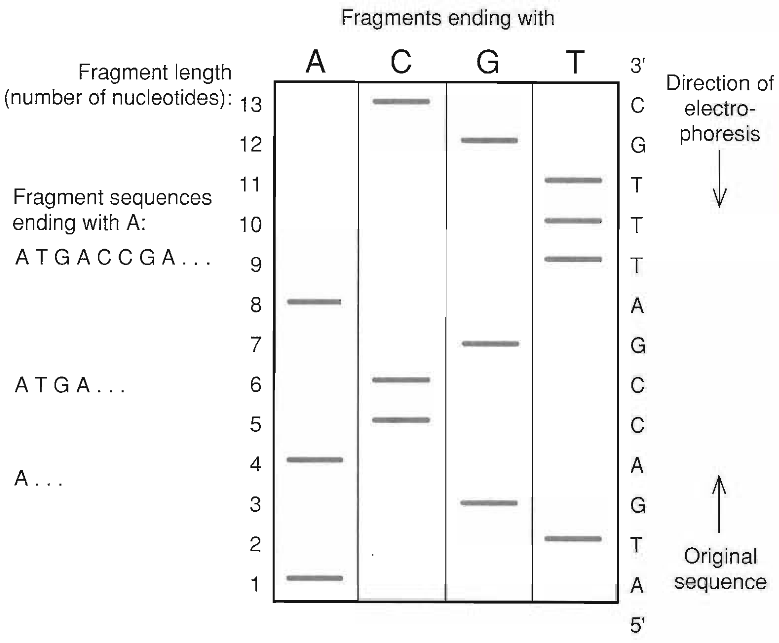
Original Strand	5'-^{32}P-ATGACCGATTTGC-3'
	5'-^{32}P-A
Labeled fragments ending in A	5'-^{32}P-ATGA
	5'-^{32}P-ATGACCGA
	5'-^{32}P-ATGAC
Labeled fragments ending in C	5'-^{32}P-ATGACC
	5'-^{32}P-ATGACCGATTTGC
	5'-^{32}P-ATG
Labeled fragments ending in G	5'-^{32}P-ATGACCG
	5'-^{32}P-ATGACCGATTTG
	5'-^{32}P-AT
Labeled fragments ending in T	5'-^{32}P-ATGACCGAT
	5'-^{32}P-ATGACCGATT
	5'-^{32}P-ATGACCGATTT

- Many copies of the strand to be sequenced are isolated and labeled with, say, the radioisotope ^{32}P, usually at the $5'$ end. The strands are chemically manipulated to create a nested set of radio-labeled fragments. By nested, we mean that each fragment in the set has a common starting point, typically at the labeled $5'$ end of the original strand, and the lengths of the labeled fragments increase stepwise, or one base at a time. In other words, the shortest fragment contains the radio label and the first base at the $5'$ end of the original strand. The next shortest fragment contains the label and the first two bases at the $5'$ end, and so on, up to the longest fragment, which is identical to the original strand.

- The fragments that make up the nested set are not prepared in one reaction mixture. Rather, copies of the original labeled strand are divided into four batches. Each batch is subjected to a different reaction, and each reaction produces labeled fragments that end in only one of the four bases A, C, T, or G. For example, if the sequence of the original labeled strand is $5'$-^{32}PATGACCGATTTGC-$3'$, the four reactions produce the four sets of labeled fragments shown in Figure 2. Together those fragments compose the complete set of nested fragments for the original strand. That is, the set includes all fragments that would be obtained by starting at the $5'$ end of the original strand and adding one base at a time.

- The fragments from the four reaction mixtures are separated by length using gel electrophoresis. A polyacrylamide gel is prepared with four parallel lanes, one for each reaction mixture. Thus each lane contains labeled fragments that end in only one of the four bases. Since polyacrylmide gels can resolve DNA molecules differing in length by just one nucleotide, the positions of all the labeled fragments can be distinguished. During electrophoresis, shorter fragments travel farther than longer fragments. Thus copies of the shortest fragment form a band farthest from the end at which the fragment batches were loaded into the gel. Successively longer fragments form bands at positions closer and closer to the loading end. Following electrophoresis, the radio-labeled fragments are visualized by exposing the gel to an x-ray filter to make an autoradiogram. Figure 3 shows the pattern of bands that would be created on the autoradiogram by the four sets of labeled fragments in Figure 2. Recall that each band contains many copies of one of those labeled fragments. The end base of those fragments is known by noting the lane in which the band appears, and the length of those fragments is determined from the vertical position of the band; fragment lengths increase from the bottom to the top of the autoradiogram. Therefore, the base sequence of the original long strand can be read directly from the autoradiogram. One starts at the bottom and looks across the four lanes to find the lane containing the band corresponding to the shortest fragments. Those fragments end at the base marked at the top of the lane. Then one continues up and across the autoradiogram, each time identifying the lane containing the band corresponding to the next longer fragments and thus identifying the end base of those fragments. The sequence of the original strand is thus read from its 5′ end, the common starting point, to its 3′ end.

Figure 3. Autoradiogram of Sequencing Gel for Simplified Example



Schematic diagram of autoradiogram showing the positions of labeled fragments generated in four reaction mixtures from the sequence 5'-^{32}P-ATGACCGATTTGC-3'. The sequence in the 5'-to-3' direction is read from the bottom to the top of the autoradiogram.

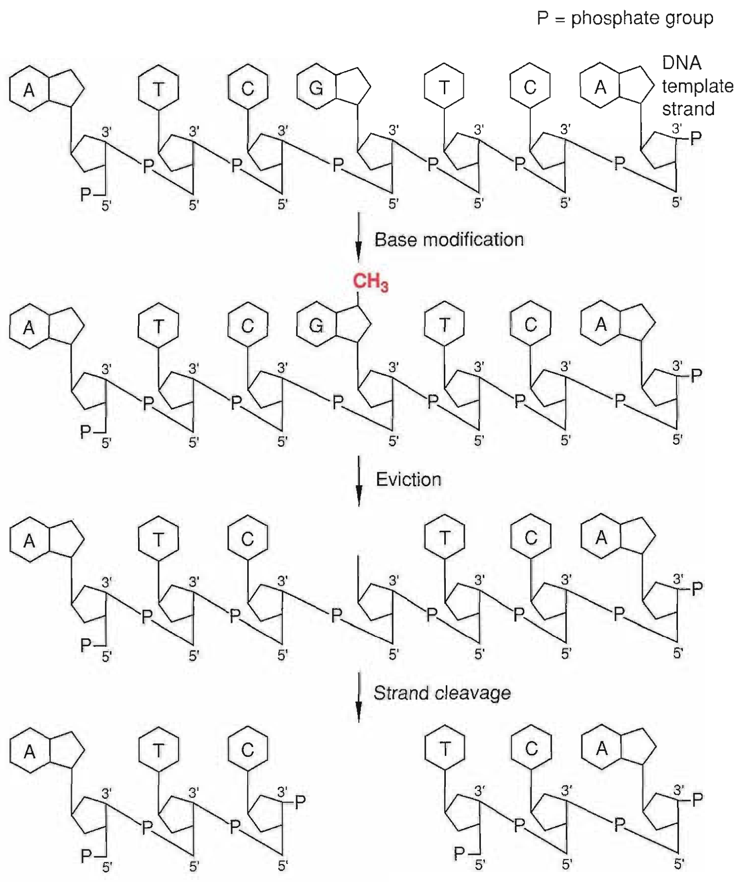
The Sanger and Maxam-Gilbert sequencing protocols differ in the reactions used to generate the four batches of labeled fragments making up the nested set. The Sanger method involves enzymatic synthesis of the radio-labeled fragments from unlabeled DNA strands. The Maxam-Gilbert method involves chemical cleavage of prelabeled DNA strands in four different ways to form the four different collections of labeled fragments. The details of the two procedures are described in Figures 4 and 5.

Figure 4. Maxam-Gilbert Sequencing Method

The Maxam-Gilbert sequencing protocol uses chemical cleavage at specific bases to generate, from pre-labeled copies of the DNA strand to be sequenced, a nested set of labeled fragments. Recall that the fragments in the set increase in length one base at a time from the 5' end of the original labeled strand. Four different cleavage reactions are used, and the reaction products are separated by length on four lanes of a gel to determine the order of the cleaved bases along the original labeled strand.

Two chemical cleavage reactions are employed; one cleaves a DNA strand at guanine (G) and adenine (A), the two purines, and the other cleaves the DNA at cytosine (C) and thymine (T), the two pyrimidines. The first reaction can be slightly modified to cleave at G only, and the second slightly modified to cleave at C only. In each reaction, cleavage of single-stranded DNA is accomplished by chemically modifying a specific base, removing the modified base from its sugar, and then breaking the bonds that hold the exposed sugar in the sugar-phosphate backbone of the DNA molecule.

(a) Cleavage Reaction for Guanine



Dimethylsulfate is used to methylate guanine. After eviction of the modified base, the exposed sugar, deoxyribose, is then removed from the backbone. Thus the strand is cleaved in two.

(b) Fragments from Single Cleavage at G

5'-^{32}P-ATGACCGATTTGC-3' —— Labeled template strand

5'-^{32}P-AT-3' 5'-ACCGATTTGC-3'
5'-^{32}P-ATGACC-3' 5'-ATTTGC-3'
5'-^{32}P-ATGACCGATT-3' 5'-C-3'

} Six different types of fragments are produced. Only three of those include the labeled 5' end of the original strand.

The reaction that cleaves guanine is shown schematically in (a). A methyl group is added to guanine, the modified base is removed from its sugar by heating, and the exposed sugar is removed from the backbone by heating in alkali. To cleave at both A and G, the procedure is identical except that a dilute acid is added after the methylation step. The reactions that cleave at C, or at C and T, involve hydrazine to remove the bases and piperidine to cleave the backbone. The extent of the reaction shown in (a) can be carefully limited so that, on average, only one G is evicted from each strand, thus each strand is cleaved at only one of its guanine sites.

A radiolabeled strand to be sequenced and the fragments created from that strand by a single cleavage at the site of G are illustrated in (b). Each original strand is broken into a labeled fragment and an unlabeled fragment. All the labeled fragments start at the 5' end of the strand and terminate at the base that precedes the site of a G along the original strand. Only the labeled fragments will be recorded once all the fragments are separated on a gel and visualized by exposing the gel to an x-ray film to create an autoradiogram of the gel.

Given the four chemical cleavage reactions, we can outline the steps involved in Maxam-Gilbert sequencing.

Step 1: Preparation of Labeled Strands. Many copies of the DNA segment to be sequenced are labeled with radioisotope ^{32}P at the 5' end of the strand. If the DNA is cloned in double-stranded form, then the 5' ends of both strands are labeled. The DNA is then denatured, copies of one strand are isolated from copies of the other strand, and each strand is sequenced separately.

Step 2: Generating a Nested Set of Labeled Fragments. Copies of one labeled strand are divided into four batches, and each batch is subjected to one of four chemical cleavage reactions outlined above. The reactions cleave the template strands at G, G and A, C, or C and T, respectively. All labeled fragments in each batch begin at the 5' end of the original strand.

Step 3: Electrophoresis and Gel Reading. The fragments from the four reactions are separated in parallel on four lanes of a gel by electrophoresis. An autoradiogram of the gel shows the positions of the labeled fragments only. A schematic of the autoradiogram is shown in the figure. Each of the four lanes is labeled by the base or bases at which the original strand was cleaved. Fragments cleaved at C show up in two lanes, the one marked C and the one marked C and T. Fragments cleaved at T are identified by noting that they appear in the lane marked C and T, but do not appear in the lane marked C. Fragments ending in A or G can be similarly identified. Note that the fragment cleaved at the first base will not show up on the gel, so the first base at the 5' end of the original strand cannot be determined. As described in the main text, the band corresponding to the shortest fragments is at the bottom of the autoradiogram. The 5'-to-3' sequence of the original strand is read by noting the positions and lanes of the bands from the bottom to the top of the autoradiogram.

(c) Steps in Maxam-Gilbert Sequencing

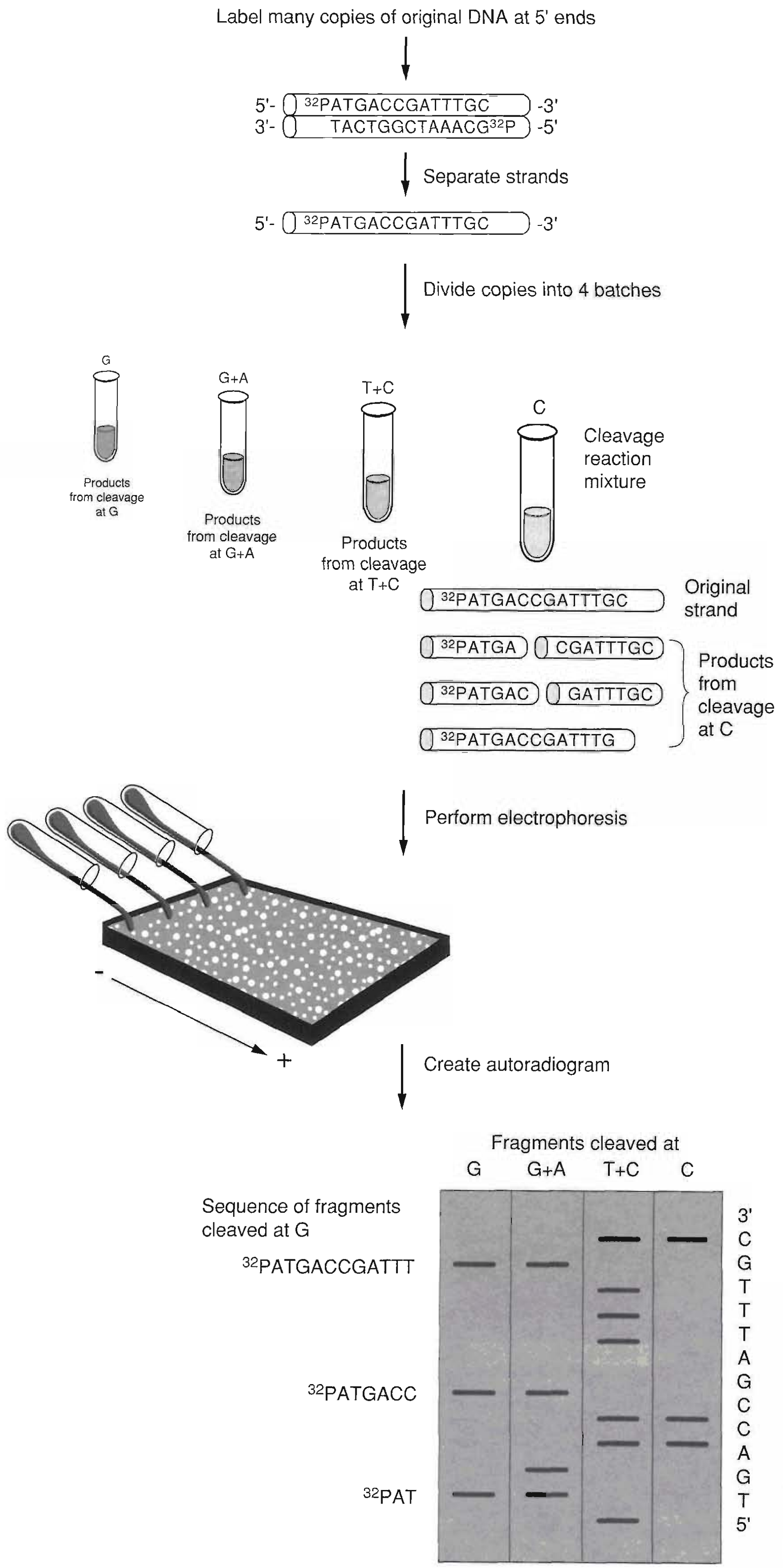
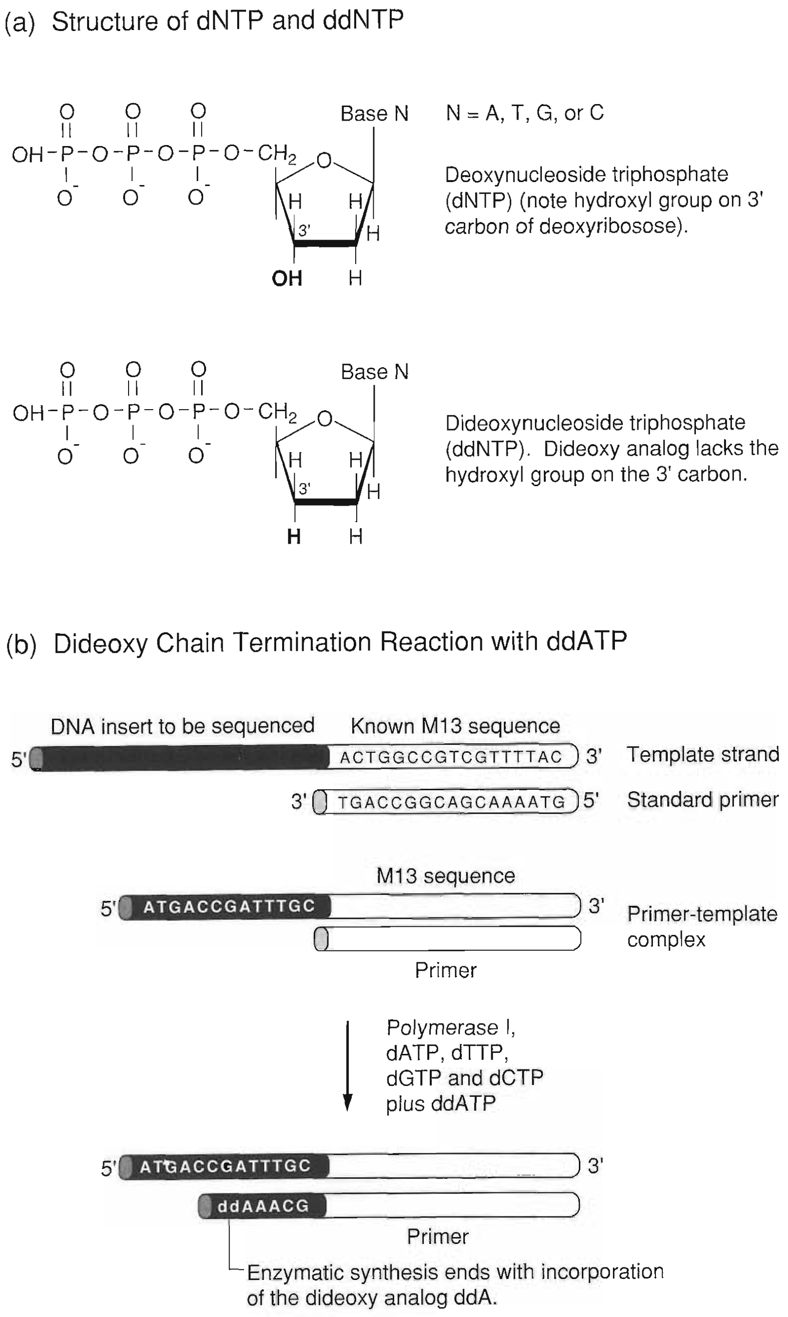


Figure 5. Sanger Sequencing Method

The Sanger method for sequencing, also known as the dideoxy chain termination method, generates the nested set of labeled fragments (see main text) from a template strand by replicating the template strand to be sequenced and interrupting the replication at one of the four bases. Four different replication reactions produce fragments that terminate in A, C, G, or T, respectively.

The replication reaction follows the path described in "DNA Replication" (see box in "Understanding Inheritance"). A DNA primer is attached (by hybridization) to the template strand and deoxynucleoside triphosphates (dNTPs) are sequentially added to the primer strand by a DNA polymerase. However, dideoxynucleoside triphosphates, say, ddATPs, are present in the reaction mixture along with the usual dNTPs. If, during replication, ddATP rather than dATP is incorporated into the growing DNA strand, then replication stops at that nucleotide.



(a) Structure of dNTP and ddNTP

N = A, T, G, or C

Deoxynucleoside triphosphate (dNTP) (note hydroxyl group on 3' carbon of deoxyribosose).

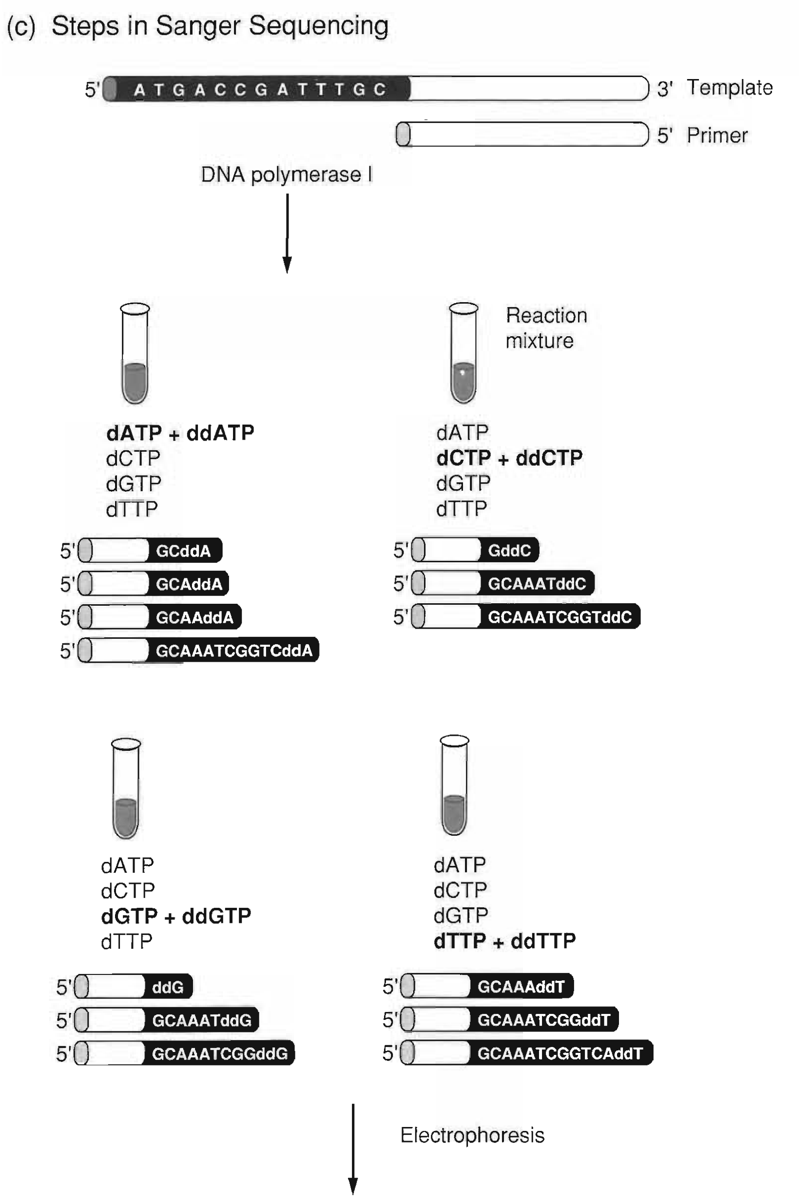
Dideoxynucleoside triphosphate (ddNTP). Dideoxy analog lacks the hydroxyl group on the 3' carbon.

(b) Dideoxy Chain Termination Reaction with ddATP

Incorporation of ddATP rather than dATP is random so all possible strands ending at ddATP are synthesized in the reaction.

In (a) we show the difference between dNTP and ddNTP. The dideoxy analog lacks the hydroxyl group that is present on the 3' carbon of the sugar in dNTP and is needed to form an O-P-O bridge to the next nucleotide. Thus, the addition of a ddNTP to the growing strand prevents the polymerase from adding additional nucleotides, and the new synthesized strand terminates with the base N. Thus all the strands synthesized in the presence of ddATP have sequences that terminate at A. These strands are complementary to the template strand, and terminate opposite the site of a T on the template strand. Complementary strands terminating in either A, G, C, or T are produced by the inclusion in the reaction mixture of ddATP, ddGTp, ddCTP, or ddTTP, respectively.

As illustrated in (b), copies of the template strand to be sequenced must be prepared with a short known sequence at the 3' end of the strand. That short sequence will then hybridize to a DNA primer whose sequence is exactly complementary to that of the known sequence. The primer is essential to initiate replication of the templates by DNA polymerase. The most convenient method for adding a known sequence to the 3' end of the template strand is to clone the strand in the single-stranded cloning vector M13 so that a known M13 sequence will always flank the unknown DNA insert and can serve as the site for binding a standard primer. Also, the M13 cloning protocol automatically creates two types of clones, each type containing a DNA insert whose sequence is complementary to that of the other DNA insert. Thus, the two complementary strands may be sequenced and the two sequences cross-checked to ensure sequence accuracy.



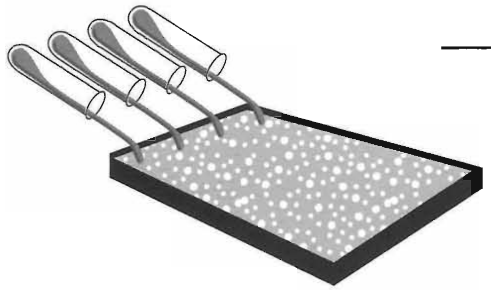
In (c) we outline the three steps involved in the Sanger dideoxy sequencing method.

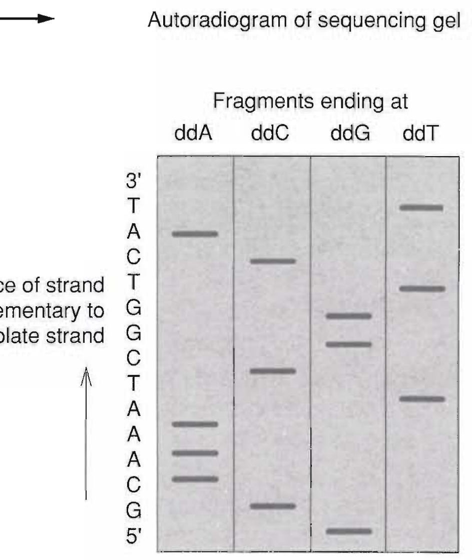
Step 1: Template Preparation. Copies of the template strand are cloned in M13. They are thus flanked at their 3' ends by a known sequence that will bind to a standard primer.

Step 2: Generating a Nested Set of Labeled Fragments. Copies of each template strand are divided into four batches, and each batch is used for a different replication reaction. Copies of the same standard primer and DNA polymerase I is used in all four reactions. To synthesize fragments, all of which terminate at A, the dideoxy analog ddATP is added to the reaction mixture along with dATP, dGTP, dCTP, dTTP the standard primer and DNA polymerase I. The ddATPs and one of the dNTPs are labeled with a radioactive isotope to produce radiolabeled strands. The figure shows a short template strand, the primer, the four reaction mixtures, and the labeled strands produced by each reaction. Note that the synthesized fragments from the four reaction mixtures compose the set of nested fragments needed to determine the order of the bases in the strand complementary to the template strand.

Step 3: Electrophoresis and Gel Reading. The fragments from the four reaction mixtures are loaded into four parallel lanes of a polyacrylamide gel and separated by length using electrophoresis.

An autoradiogram of the gel is read as described in the main text to determine the order of the bases in the strand complementary to that of the template strand. Again, since the bands corresponding to the shortest fragments are at the bottom of the autoradiogram, the 5'-to-3' sequence of the strand complementary to the template strand is read from the bottom to the top of the autoradiogram.




Autoradiogram of sequencing gel
Fragments ending at
ddA ddC ddG ddT
3' T A C T G G C T A A A C G 5'

Sequence of strand complementary to template strand

The final step in both procedures is to separate the labeled fragments by length using gel electrophoresis (see "Gel Electrophoresis" in "Understanding Inheritance"). Since the fragment mobility in the gel varies as the reciprocal of the logarithm of the fragment length, shorter fragments are more widely separated from one another than longer fragments. That is, the resolution of fragment lengths decreases as the fragment length increases. Therefore, the range of fragment lengths that can be resolved in a single gel is limited to several hundred bases. Moreover, the separation of fragments in a standard gel (0.2 to 0.4 millimeters thick) is a relatively slow process. At least several hours are required to resolve fragment lengths from one to several hundred bases long. [More recently, very narrow gel-filled capillary tubes have been used to decrease the time needed for fragment separation. Several hundred bases can be resolved in tens of minutes and the resolution is high enough to read 1000 bases from a single gel.] The average error rate in a single sequencing run is about 1 base in 100. The errors are often due to inhomogeneities in the gel and various sequence-dependent conformational changes in the single-stranded fragments that affect their mobility in the gel.

Since only short stretches of DNA, several hundred to a thousand base pairs in length, can be obtained from a single sequencing gel, many short sequences must be generated separately and then combined to determine the sequence of a much longer DNA fragment. Various strategies have been developed to generate these short sequences from the larger fragment.

The "shotgun" approach is the most widely used in the larger sequencing projects. Copies of a long fragment to be sequenced are broken into much shorter fragments that overlap one another, and the short fragments are cloned. Those clones are then picked at random and sequenced. The sequence of the long fragment is determined by finding overlaps among the short sequences and assembling those sequences into the most likely order. Numerous computer algorithms have been developed to facilitate the assembly of long sequences.

Inevitably, gaps remain in the sequence of the long fragment, and they are filled by switching to a directed sequencing strategy. That is, the short clones are no longer sequenced at random, but rather, short sequences at the end of a continuous stretch of known sequence provide the information necessary to construct a probe to pick out a clone, or region of a clone, whose sequence will extend the known sequence. Most of the large sequencing projects to date have used a mixture of random and directed sequencing strategies to complete the sequence of long, contiguous stretches of DNA. The advantage of the random, or "shotgun," strategy is that in the course of picking clones at random and sequencing them, any given region is usually sequenced many times, thereby reducing the errors in the final sequence.

Almost all steps involved in sequencing are amenable to automation, and through automation many groups hope to increase both the throughput and the consistency of large-scale sequencing efforts. Several automatic sequencing machines have been

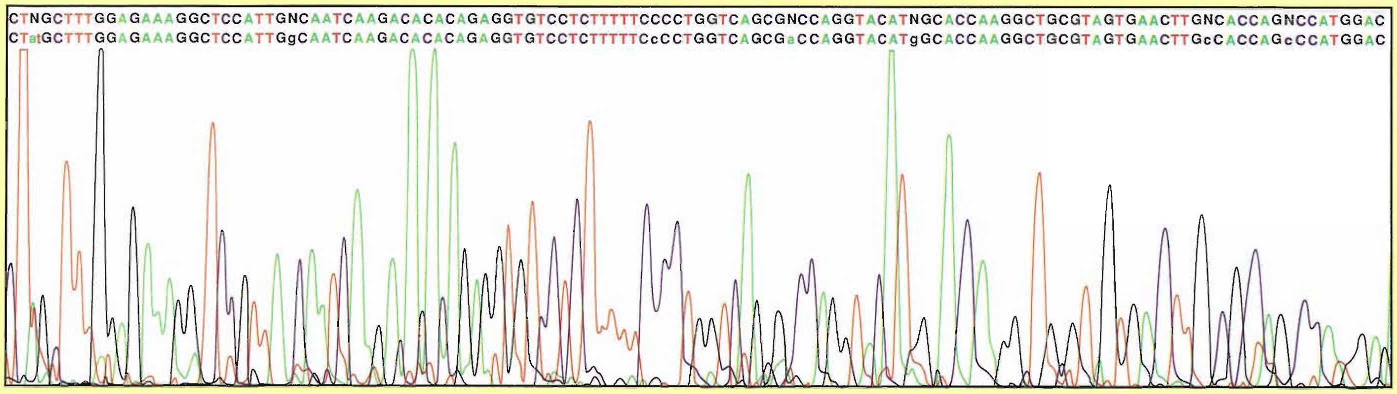


Figure 6. Output of Automatic Sequencing Machine
Each of four dideoxy sequencing reactions produces fragments labeled with a dye that fluoresces at a different wavelength. As the fragments from the four reactions migrate down a single lane of a polyacrylamide gel, they pass through a laser beam and produce a fluorescence signal. The machine automatically records the signal and calls the end base of the fragments based on the color (wavelength) of the fluorescence signal. The sequence of the strand complementary to the template strand is read from right to left corresponding to the 5′-to-3′ direction. The machine automatically generates the top sequence, recording any ambiguity in the base call as an N. A technician can resolve most such ambiguities by direct examination of the fluorescence signals. If the technician concludes with high certainty that a particular N is, for example, the base G, he or she replaces that N with a g in the bottom sequence.

on the market for a number of years. Those machines automate the steps of gel electrophoresis, gel reading, and the "calling" of the end bases of the successively longer fragments. The machines designed for high throughput require that the fragments produced by the four sequencing reactions be labeled with fluorescent dyes rather than radioisotopes, and they employ laser-induced fluorescence to detect the order of the labeled fragments as they migrate through the gel. Some machines use four parallel lanes for the fragments of the four reaction mixtures; others use a single gel lane for all the fragments. The output of a high-throughput sequencing machine includes a plot of the fluorescence signals versus time produced as the fragments migrate past the laser as well as the sequence of bases corresponding to the time sequence of the variously colored fluorescence peaks. Ambiguities in the data are also noted automatically (see Figure 6).

Under optimal conditions, the automatic sequencers are capable of producing 12,000 base pairs of raw data per day. However, much work remains to improve reliability and to organize the efficient use of those machines in large-scale sequencing projects. For example, problems associated with the preparation of clones for sequencing, the checking of the short sequences and assembling them into longer contiguous sequences, and the tracking of all procedures involved in sequencing need increased attention. So far, despite the availability of automatic sequencing machines, production of finished sequence remains a slow and expensive process. Those working on improving existing technologies and streamlining their use expect to achieve a tenfold increase in sequencing throughput within the next few years, and perhaps a hundredfold increase in ten years. Others are involved in developing radically new sequencing technologies that, if successful, might achieve the hundredfold to thousandfold increase needed to sequence the entire human genome. (See the discussion of new technologies in Part III of "Mapping the Genome" as well as "Rapid DNA Sequencing Based on Single Molecule Detection.") ■

Further Reading

T. Hunkapiller, R.J. Kaiser, B.F. Koop, L. Hood "Large-Scale and Automated DNA Sequence Determination." *Science*, October 4, 1991.

Informatics—information handling and analysis

Bob Moyzis: We've been talking about improving technology to generate data much faster than we're now doing, and that brings up the problem of how to store, analyze, and distribute the data to the community. Even at the present rate, the genome centers have run smack into the issue of information handling.

David Galas: I want to emphasize that the principal resource to come from the Genome Project is an ongoing public database of information about chromosomes, segments of chromosomes, genes, and so on. So, even in this relatively early stage of the Project, we are focused on trying to envision that database and on organizing the information already available.

David Botstein: There are a lot of database types who are thinking about this problem, but at this point we don't have enough data to formulate the problem properly. Fully integrated databases for organisms don't really exist yet. In the long run, creating those databases is going to be a major problem. The Genome Project has established a joint informatics task force to address the problem, but it's a very contentious group. The one thing they agree on is that the database must be useful to biologists.

David Galas: In talking to biologists, computer scientists, and mathematicians, it's clear no one has a very good concept for the ultimate database. It is also clear that we must start with some kind of database and then set up a process by which it can evolve to meet future need.

In a short term, the next couple of years, the genome database at Johns Hopkins is going to be our database, because it has no competitors. It has all the genetic data that people are willing to put into the public domain, and plans are now being laid for including physical-mapping data. That database is well conceived in present technology, and although it's clearly not at the cutting edge of database technologies, we have to do something now; we can't afford to wait. In the long term, either it will evolve into something quite different, or it will be replaced by something else.

The principal resource to come from the Genome Project is an ongoing public database of information about chromosomes, segments of chromosomes, genes, and so on. So, even in this relatively early stage of the Project, we are focused on trying to envision that database and on organizing the information already available.

Bob Moyzis: I think there is an incredible amount of data that is currently inaccessible in public databases, for example, the data sitting in DOE and NIH genome centers. Just trying to get useful genetic-mapping data out of GDB, the Genome DataBase as Johns Hopkins, is a frustrating task. They're working hard at improving this resource, but it's still an enormous task. Until data flow from the genome centers to GDB is more efficient and until GDB becomes a more user-friendly database, I'm afraid much of the information will remain in local databases.

David Botstein: The more sophisticated computer-types think that major improvements in database structure are in the pipeline, so it's clear that we shouldn't lock ourselves in. Almost everybody believes that databases currently used by people who are not computer-science experts are going to have problems. And since the new methods put additional constraints and also additional liberties on how you do things, we must get everybody to preserve their data in such a way that they don't lose any essential parts.

Norton Zinder: We're trying very hard to put together a task force to look at this problem in a very serious way. We're also making minimal databases, so that people can get the data they want quickly and not get lost in the mountains of extraneous information that are presently being stored.

Nancy Wexler: We now have many collaborations organized around cells, parts of chromosomes, and disease genes, and they are forcing people to create databases. For example, seven different laboratories around the world collaborate on Huntington's disease. They're trying to figure out their own collaborative databases and communication systems, and people are getting locked into particular formats, so the database problem needs to be addressed before it becomes unmanageable.

David Galas: I think this worry of being locked into particular data structures and so forth is a red herring. There will always be a need to redo things, to turn over equipment, and so on. That is an ongoing cost of any database. But it is a misconception to think that choosing one format locks you in forever.

Software technology is now reaching the point where you can change from relational databases to the newer object-oriented databases. The change is not trivial, but you don't have to redo everything. Biologists are afraid, as David Botstein often says, of the Stalinism of setting standards, and they use that excuse to argue against doing anything. But we desperately need to do something now because people, particularly in the smaller labs, have to be able to have access to the data.

In some ways, David's argument that we need to wait cuts against the philosophy he espouses. That is, by not doing something now we cut out all the small labs. A small university is not going to have access to anything if there is no database. So it's a real problem. But the problem is not long for this world because the education of smart people like David and the biological community in general is going to come along rapidly.

We need a great deal more communication and coordination in the informatics area. The database issue is critical now. It is an administrative problem, a software problem, a networking problem, and a research problem. It's a mess and it needs to be addressed because data is our ultimate product.

Norton Zinder: Some informatics people want to completely restructure relational databases to apply, generically, to any world and to any problem. But we have some finite problems that need immediate solutions.

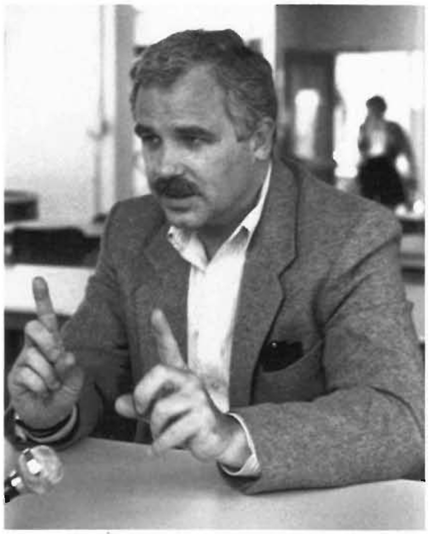
David Galas: The handling of mapping data is one such problem, and the national labs recognized the need for sophisticated data handling a long time ago. Now, as Bob mentioned, the NIH centers are beginning to recognize the problem because they're having trouble dealing with all the data.

Bob Moyzis: They're beginning to realize they're all underfunded because the money they asked for is to do the biology and there's nothing left to do all of the other things.

David Galas: Before these centers really got started, the people involved were saying things quite antithetical to what they're now saying. So clearly there's a great deal of education in the community that needs to be done. We have a bit of a two-cultures problem. On one side you have those in mathematics and computational biology and on the other side are those with a classical biology background who are doing the good work in genome mapping. So we need informatics support for the mapping efforts.

Bob Moyzis: Other than STSs, for which it is easy to construct a database and share that information, management of mapping data is very difficult. At Los Alamos we have accumulated more information on chromosome 16 than we ourselves can access in an easy fashion. It has turned out to be a bottleneck for us. The problem of sending 4,000 clones someplace is easy compared with sending the information we've accumulated on chromosome 16 in some useful and intelligible format.

On the other hand, progress on map integration, analysis, and display under



David Galas

Biologists are afraid . . . of the Stalinism of setting standards, and they often use that excuse to argue against doing anything. But we desperately need to do something now because people, particularly in the smaller labs, have to be able to have access to the data.

Bob Moyzis

If you have a bad fit between person and problem, it's frustrating all around. Certain aspects of this project are moving so fast that the informatics types need to come up with a quick and dirty solution to be of help.

the direction of Jim Fickett and others at Los Alamos have progressed to the point that I will make a prediction. Individual genome centers *must* consolidate their own data, or they will not produce a quality map. The idea that some central database, like GDB, can provide this function is nonsense. The central database should use the GenBank model. The *investigators* will produce the map. The central database will make it *available* in some consistent form.

David Galas: Another important area in informatics is research into future algorithms. We need new algorithms for doing pattern recognition in sequence data, for finding the genes among the sequence of bases, for finding similarities among sequences, and for assembling long stretches of sequence from short stretches.

Lee Hood: The information problems are tough. There are no programs that can search through a DNA sequence and unequivocally pick out the coding sequences. Scientists at Oak Ridge have made striking contributions to solving this problem. I think we're attracting good people into this field, people who know some biology. But there's really an enormous amount of work to be done. Scientists such as Chris Fields are incorporating various features of genes into their search algorithms, such as statistical asymmetries among groups of three bases or six bases in the protein-coding regions, properties of RNA splicing points and splicing boundaries, and so on. But we need to accumulate more sequence data and learn a lot more about those features before we'll have reliable algorithms for finding genes directly from the nucleotide sequence.

David Botstein: Another difficult problem is to figure out when one should be impressed with the similarity between two sequences. A related problem is to find the similarities among sequences in a huge mass of data, which means lots of pairwise comparisons. Parallel computing is very appropriate for this task of making many comparisons of many sequences, and aligning them optimally. There are a few major mathematicians who work on this problem. The most prominent are probably Sam Karlen and Michael Waterman. The business is largely combinatorics.

David Galas: I understand from my mathematical colleagues that the mapping and sequencing data present some very important and interesting mathematical problems. And often those problems that the biologists think are trivial are really difficult for the mathematicians to solve rigorously and vice versa. For example, biologists have assembled physical maps from restriction fragment lengths, but to do it in a rigorous fashion turns out to be an NP-complete problem—which means that the number of computations required to do the problem increases exponentially with the number of fragments.

Bob Moyzis: Lander and Waterman took a rough cut at that problem, but to do it in a rigorous probabilistic sense is very tough indeed. David Torney at Los Alamos is working on this problem.

David Galas: Sequence-matching problems are also very difficult if the parameters are set in a sufficiently loose way, that is, if you allow gaps and insertions in the sequences to be matched. The difficulty of the comparison is dependent on the parameters in a way that's unexpected—at least to most biologists. The sequence-assembly problem is of the same kind—NP-complete, much like the traveling-salesman problem. An approximate solution is not too hard to get if you don't have too many short sequences to assemble into one. But when you're doing massive sequencing and trying to assemble the pieces in a rigorous and automated way, then you have to worry about the real nature of the problem. Other things, like quickly searching the database for particular sequences, sound difficult to a lot of biologists, but in fact database searches are not too hard. And many biologists can do it on their computers right now because the software is available.

We're really experiencing a blossoming of this interface between biology, mathematics, and computation. That interface holds a great deal of the future of biology, much like the automation problem in engineering.

We've talked about public databases, lab support, and research on algorithms, but you can't always distinguish one of them from the others. The lab biologist trying to do the mapping problem would say, "Give me some computer guys so I can do X." The computer guy will work on that for a while and say, "Okay, now you can do X." Then the lab guy says, "While you were fixing it so I could do X, I changed my mind. We've got this new technique, now I want to do Y." Developing software and techniques for ongoing, evolving technologies is a real problem.

Nonetheless, I want to make a distinction between developing software, showing people how to use it, and making it bulletproof, versus solving the much more abstract and esoteric problems. If done by the right people, the abstract problems are going to be extremely important. It's almost another two-cultures problem. Biologists say they want the mathematicians to be their computer programmers, but we also need to tackle the difficult mathematical problems.

Bob Moyzis: Very different kinds of people are interested in doing those very different types of problems. If you have a bad fit between person and problem, it's frustrating all around. Certain aspects of this project are moving so fast that the informatics types need to come up with a quick and dirty solution to be of help.

On the other hand, there are some major problems that aren't going to go away in two years and need more long-range kind of work. A few years ago many mathematical types were trying to develop models of the mapping problems, hoping to find the best strategy for mapping the genome. But molecular biologists, at least the more aggressive ones, aren't willing to wait around for anything. They want to get the job done on this project, and they'll switch midstream if a new technique comes on-line that looks better.

The *better* technique is very difficult to define because it depends on personal preference and skill at certain techniques not simply on some abstract measure of efficiency. Simulations of the mapping problems show only slight differences in the efficiency of different strategies and are really not that informative. So the mathematical problems have to be chosen with some care.

David Botstein: The recurrent dilemma that we haven't touched on at all is who will have access to the data. Most scientists want the data to be in a public database as soon as you read the sequences off your gel. But the sooner one releases the data, the less chance one has to check the data. It's a trade-off between speed and accuracy. There are also commercial and patent concerns because the sequences have many biotechnology spin-offs. That's a very difficult and touchy subject.

Most scientists want the data to be in a public database as soon as you read the sequences off your gel. But the sooner one releases the data, the less chance one has to check the data. It's a trade-off between speed and accuracy. There are also commercial and patent concerns because the sequences have many biotechnology spin-offs. That's a very difficult and touchy subject.